Interim Report

(1)

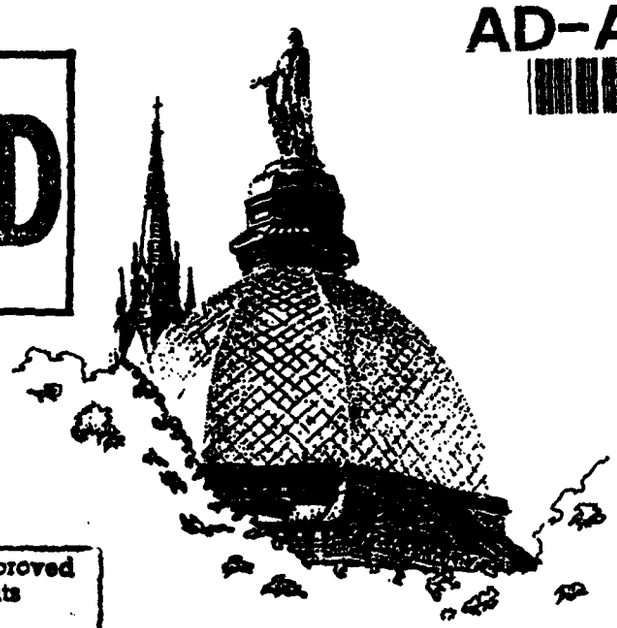
High-Speed Fixed and Floating Point Implementation of Delta-Operator Formulated Discrete Time Systems

work performed for:

The Office of Naval Research

AD-A283 109

DTIC
SELECTED
AUG 17 1994
F



This document has been approved for public release and sale; its distribution is unlimited.

January 1, 1994 - June 30, 1994

Principal Investigator:
Professor Peter H. Bauer

94 8 08 092

Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556

94-24995
2 pgs

DTIC QUALITY INSPECTED 1

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	08/05/94	Interim 01/01/94 - 06/30/94

4. TITLE AND SUBTITLE
High-Speed Fixed and Floating Point Implementation of Delta-Operator Formulated Discrete Time Systems

5. FUNDING NUMBERS
Grant #:
N00014-94-1-0387
Project Code:
3148509-01

6. AUTHOR(S)
Peter H. Bauer

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dept. of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Code 251 : Jwk
Ballston Tower One
800 N. Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Report was prepared in cooperation with Prof. K. Premaratne, Dept. of Electrical & Computer Engr., Univ. of Miami, Coral Gables, FL 33124

12a. DISTRIBUTION/AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words) This report addresses the analysis and design of finite wordlength implementations of linear t.i.v. delta-systems and the development of a 2-D delta-operator state space model. It is shown that in fixed point arithmetic linear t.i.v. systems implementation in delta-operator form do not generally outperform their q-operator counterpart. In fact, delta-operator systems always show unstable limit cycle behavior and convergence to incorrect equilibrium points, independent of the choice of the realization or the sampling time. The coefficient sensitivity for delta-systems is still superior to the shift-operator. In the case of floating point arithmetic, delta-operator implementations perform consistently better than their shift-operator counterparts. Delta-systems show superior quantization noise and sensitivity properties. The zero-convergence problem of the fixed point case does not exist if the mantissa length is chosen sufficiently large. Due to its attractive finite wordlength properties, the concept of delta-operators has been extended to the multi-dimensional case. A 2-D state space model was developed and the notions of reachability and observability gramian and balanced realization have been introduced. The problem of directly checking stability in the delta-domain has also been addressed. Similarly to the 1-D case, the sensitivity & roundoff noise behavior was analyzed.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Finite Wordlength, Stability, Quantization Errors, Limit Cycles, Sensitivity, Fixed and Floating Point Arithmetic, Multi-Dimensional Systems	10
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRAC
Unclassified	Unclassified	Unclassified	UL

Table of Contents

Accesion For

NTIS CRA&I ☒

DTIC TAB ☐

Unannounced ☐

Justification .....

By .....

Distribution /

Availability Codes

Dist	Avail and / or Special
A-1	

SEMIANNUAL PERFORMANCE REPORT
GRANT NO's: N00014-94-1-0387

Summary of Phase P1 Results

Phase P1 consists of two tasks:

[T1] Task T1: Analysis and design of finite wordlength implementations of linear, time-invariant δ-Systems.

[T3] Task T3: 2-D and m-D δ-system models.

The major part of task T1 was carried out at the University of Notre Dame by Dr. Peter H. Bauer while the major part of task T3 was carried out at the University of Miami by Dr. Kamal Premaratne under grant No. N00014-94-1-0454. The project being an extensive collaborative effort, the two PI's have been in constant contact.

The following is a summary of the phase P1 results.

Task T1: Analysis and Design of Finite Wordlength Implementations of Linear, Time-Invariant δ-Systems

The conclusions drawn from the work conducted for task T1 may be summarized as follows:

1. The Fixed-Point Arithmetic Case: When limit cycle performance is crucial, the q-operator implementation is preferrable. The δ-operator implementation is superior with regard to coefficient sensitivity issues.

2. The Floating-Point Arithmetic Case: Generally, the δ-operator implementation outperforms its q-operator counterpart. In particular, in high-order and high-speed applications, the δ-operator implementation is the best choice.

Prior to a more detailed exposition, first we provide qualitative justification for the above conclusion. The state equations of a δ-operator system can be written as:

$$\begin{aligned} \delta[\mathbf{x}](n) &= A_\delta \mathbf{x}(n) + B_\delta \mathbf{u}(n); \\ q[\mathbf{x}](n) &= \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n). \end{aligned} \tag{T1.1}$$

where $\mathbf{x}$ and $\mathbf{u}$ are the state and input vectors, respectively. Here, Δ denote a positive real constant (typically, the sampling time). The symbol $\delta[\cdot]$ denotes the δ-operator, that is,

$$\delta[\mathbf{x}](n) = \frac{q[\mathbf{x}](n) - \mathbf{x}(n)}{\Delta} = \frac{q-1}{\Delta}\mathbf{x}(n), \tag{T1.2}$$

and $q[\cdot]$ denotes the usual q-operator, that is,

$$q[\mathbf{x}](n) = \mathbf{x}(n+1). \tag{T1.3}$$

The corresponding formulation of (T1.1) in terms of the q-operator is

$$q[\mathbf{x}](n) = A_q\mathbf{x}(n) + B_q\mathbf{u}(n), \tag{T1.4}$$

where

$$A_q = I + \Delta \cdot A_\delta \Longleftrightarrow A_\delta = \frac{A_q - I}{\Delta} \quad \text{and} \quad B_q = \Delta \cdot B_\delta \Longleftrightarrow B_\delta = \frac{B_q}{\Delta}. \tag{T1.5}$$

Now, given $\mathbf{x}$ and $\mathbf{u}$, both representations compute $q[\mathbf{x}]$ with a certain accuracy. Consider the δ-operator formulation in (T1.1). Here we encounter two errors:

1. The first is due to the computation of $\delta[\mathbf{x}]$, that is, the first equation in (T1.1). We will refer to this equation as the *intermediate equation.*

2. The second is due to the eventual computation of $q[\mathbf{x}]$, that is, the second equation in (T1.1). We will refer to this equation as the *update equation.*

Let us assume that the total error in computing $q[\mathbf{x}]$ is mainly due to the intermediate equation in (T1.1) (rather than the update equation). Then, by choosing Δ sufficiently small, the total error in computing $q[\mathbf{x}]$ will be approximately the error created by the update equation which is small!. In this case, the δ-operator representation has better finite wordlength properties than its q-operator counterpart in (T1.4).

If, however, the errors accumulated in the intermediate and the update equations in (T1.1) are comparable, $q[\mathbf{x}]$ computed through the δ-operator representation will show approximately the same error as that computed through its q-operator counterpart assuming Δ is sufficiently small. If Δ is not sufficiently smaller than one, the δ-operator representation will actually perform worse than the q-operator representation!

If the error introduced in the update equation is larger than that in the intermediate equation, the δ-operator representation would consistently perform worse!! In reality, this case is very unlikely to occur.

Next, a more detailed exposition follows.

T1.1 The Fixed-Point Arithmetic Case

We now discuss some of the results regarding the fixed-point (FXP) case. Here, our results

in fact indicate that, in case limit cycle behavior is crucial, the δ-operator representation is NOT suitable with this arithmetic scheme [1]. Such a case may occur when nonlinear systems are implemented through FXP δ-operator based schemes.

Zero-input limit cycles. Independent of Δ, zero-input limit cycles cannot be avoided in FXP δ-implementations. This is easily explained as follows: If Δ is chosen very small, the contribution from the intermediate equation being small (since $\delta[\mathbf{x}]$ is being multiplied by Δ), during the update equation, $q[\mathbf{x}]$ can be quantized to $\mathbf{x}$ creating a DC limit cycle, that is, an incorrect equilibrium point different from zero results. We emphasize that, most of the desirable properties of δ-operator implementations are based on a small Δ. We may also show that, if Δ is chosen larger (this case is of course somewhat less important), DC limit cycles will still exist. Hence, δ-operator representations cannot be implemented limit cycle free in FXP format! This fact is independent of the particular realization of the system.

Deadband size. Since δ-systems cannot be implemented limit cycle free in FXP format, it is of interest to investigate te the size of such limit cycles since, in certain situations, such small limit cycle amplitudes can be tolerated. It can be shown that, the magnitude of Δ determines the magnitude of the limit cycle. The smaller the Δ, the larger will be the deadband and hence the limit cycle magnitude. An approximate relationship regarding this is

$$\Delta \times \text{size of deadband} = 1, \tag{T1.6}$$

where the size of deadband is measured in multiples of the quantization step size. Here, the deadband corresponds to that obtained by considering the quantization of $\Delta \cdot \delta[\mathbf{x}]$. Therefore, the usual choice of a small Δ creates a larger deadband!

The input driven case. Although the input driven case is not part of the originally proposed work, some interesting results have been obtained. For small values of Δ, there exists a bounded input signal that does not allow control of the state trajectory. In other words, given sufficiently small Δ, the state trajectory may not be influenced by such an input signal.

The influence of the realization. First, it was necessary to develop a suitable scheme to investigate the effect of realization on the presence or absence of limit cycles. In this direction, for the q-operator case, a computer-based exhaustive search algorithm that checks for limit cycles (DC and/or oscillatory) has been developed [5].

As discussed before, we have shown that, a stable linear time-invariant δ-system cannot be implemented limit cycle free in FXP. The size of the deadband however also depends on the particular realization, that is, the structure of A_δ. Given a system transfer function, there are forms which minimize this deadband size with respect to some appropriately chosen measure. For example, in order to minimize DC limit cycle amplitude, one may choose the normal form (in terms of A_δ) as a suitable candidate.

The influence of quantization nonlinearity and its deadzone. Since a larger deadzone implies larger DC limit cycle amplitudes, the use of quantizers with reduced, or even zero, deadzone was therefore proposed. In investigating first-order systems, by reducing the deadzone, it was found that, existence of DC limit cycles can indeed be reduced. Unfortunately, other oscillatory limit cycles will be created. This phenomenon is due to the increased gain exhibited towards small input signals by the quantizer.

Scaling. As discussed above, we have shown that, independent of either the form of A_δ or the magnitude of Δ, a FXP implemented δ-system cannot be free of zero-input limit cycles. Hence, scaling cannot be offered as a possible solution.

T1.2 The Floating-Point Arithmetic Case

The floating-point (FLP) implementation of δ-systems is currently under investigation. The results obtained so far are very encouraging, and indicate that, quantization errors due to FLP arithmetic have a much smaller effect on the system behavior than in the FXP case. In fact, preliminary results show that, for δ-systems of order three and higher, errors in computing $q[\mathbf{x}]$ can be made significantly smaller than for the corresponding q-systems. This is because, for a FLP implementation of such a system, errors created through the intermediate equation are larger than those created through the update equation. As previously mentioned, in this situation, δ-systems behave better than their q-operator counterparts!

Limit cycles. In FLP arithmetic, a linearly stable time invariant system, under zero-input conditions, may exhibit four types of responses: A diverging response, an oscillatory periodic response of arbitrary magnitude, an oscillatory periodic response in underflow, or an asymptotically stable response. Only the last two response types are acceptable in practice. It is well known that, the last response type is in fact a very stringent requirement and is often not required in practice. Results so far obtained show that, when the requirements for a response in underflow are compared, the δ-system requires less wordlength than its q-system counterpart! This advantage in fact grows with the order of the system!!

Once the system reaches underflow conditions, the δ-system again exhibits DC limit cycles. However, if the exponent register is chosen sufficiently large, the amplitude of these oscillations can be made extremely small and hence, for all practical purposes, this problem is solved.

Deadband size. If the condition on the mantissa length that guarantees convergence into underflow is satisfied, then the deadband size will be very small. Hence, it can be neglected for all practical purposes. This assumes a properly chosen exponent register length since the exponent register length determines the dynamic range of underflow.

The Influence of the Nonlinearity. Unlike the FXP case, the characteristic of the nonlinearity has only a minor effect on the system behavior, significant differences being present only in underflow conditions

The Underflow case. In underflow, the δ-system seems to behave worse than its q-operator counterpart. This is mainly due to the fact that, a FLP system in underflow essentially performs very similar to a FXP system. However, as mentioned above, if the dynamic range of underflow is chosen properly, the system behavior in underflow is of little practical interest.

Block Floating-Point Arithmetic. Even for the q-operator case, results regarding block FLP implementations are lacking. Hence, investigations regarding block FLP implementation of δ-systems is in its early stages. In order to obtain a comparison between the two types of implementations, current research is geared towards obtaining results applicable for the q-operator case.

T1.3 The Multi-Dimensional Case

The results on one-dimensional (1-D) δ-operator implementations in FXP arithmetic directly carry over to the multi-dimensional (m-D) case. The existence of non-converging responses along the boundary of the causality region can easily be proven using the same type of argument used in the 1-D case. Consequently, δ-operator based implementations of m-D systems cannot be implemented limit cycle free in FXP.

Task T3: 2-D and m-D δ-system models

Discrete-time systems implemented using the δ-operator, as is clear from the discussion above, exhibit superior finite wordlength properties with FLP arithmetic. In the case of FXP arithmetic, they still provide superior coefficient sensitivity. The development of 2-D and m-D models applicable for δ-operator implementations was hence motivated with the

expectation that these properties would still hold true.

The conclusions drawn from the work conducted for task T3 may be summarized as follows: Similar to the 1-D case, under FLP arithmetic, the δ-operator implementation of 2-D and m-D discrete-time systems provides the best choice. Again, this is particularly true in high-order and high-speed applications.

State-space models. In Roesser local s.s. model of q-operator formulated 2-D discrete-time systems takes the form

$$\begin{aligned}\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A_q^{(1)} & A_q^{(2)} \\ A_q^{(3)} & A_q^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B_q^{(1)} \\ B_q^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\ &\doteq [A_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B_q]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &= \begin{bmatrix} C_q^{(1)} & C_q^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j) \\ &\doteq [C_q] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D_q]\mathbf{u}(i,j), \end{aligned} \tag{T3.1}$$

where $A_q^{(1)}$ is of size $n_h \times n_h$, $A_q^{(4)}$ is of size $n_v \times n_v$, etc. Also, $q_h[\cdot]$ and $q_v[\cdot]$ denote the horizontal and vertical shift operators, that is,

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j) \quad \text{and} \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{T3.2}$$

To exploit the advantages of δ-operator implementations, analogous to the 1-D case, we define the operators

$$\begin{aligned} \delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\Delta_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_h}; \\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\Delta_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\Delta_v}, \end{aligned} \tag{T3.3}$$

where Δ_h and Δ_v are two positive real constants. The corresponding δ-operator s.s. model may then be obtained as

$$\begin{aligned}\begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} &= \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix} \mathbf{u}(i,j) \\ &\doteq [A] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &= \begin{bmatrix} C^{(1)} & C^{(2)} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j) \\ &\doteq [C] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j). \end{aligned} \tag{T3.4}$$

This is the 2-D version of the intermediate equation mentioned earlier. In addition, as for the 1-D case, we have the following update equations:

$$\begin{aligned} q_h[\mathbf{x}^h](i,j) &= \mathbf{x}^h(i,j) + \Delta_h \cdot \delta_h[\mathbf{x}^h](i,j); \\ q_v[\mathbf{x}^v](i,j) &= \mathbf{x}^v(i,j) + \Delta_v \cdot \delta_v[\mathbf{x}^v](i,j). \end{aligned} \tag{T3.5}$$

Note that,

$$\begin{aligned} A_q = I + \boldsymbol{\Delta} \cdot A_\delta &\iff A_\delta = \boldsymbol{\Delta}^{-1} \cdot (A_q - I_n); \\ B_q = \boldsymbol{\Delta} \cdot B &\iff B_\delta = \boldsymbol{\Delta}^{-1} \cdot B_q; \\ C_q = C_\delta &\iff C_\delta = C_q; \\ D_q = D_\delta &\iff D_\delta = D_q. \end{aligned} \tag{T3.6}$$

Here, $\boldsymbol{\Delta} = [\Delta_h I_{n_h} \oplus \Delta_v I_{n_v}]$ is of size $(n_h + n_v) \times (n_h + n_v)$.

The associated system theoretic notions, such as, transition matrix, transfer function, characteristic equation, etc., have also been introduced. This s.s. model is the basis for designing 2-D filters with superior finite wordlength properties. The design procedures developed are expected to be extremely useful in obtaining high-Q 2-D and m-D digital filters that are suitable for high-speed applications.

Stability. In the 1-D case, it has been shown that, direct techniques with no recourse to transformations (that first converts a given δ-system to its q-system counterpart) can provide numerically more reliable stability checking algorithms. With this in mind, for the 2-D case, a direct stability checking technique applicable to the corresponding δ-system transfer function has been introduced. For this purpose, a recently developed tabular form was extended to the complex coefficient case and the notion of Schur-Cohn minors was introduced to the δ-operator case.

Gramians and balanced realization. The notions of reachability and observability gramians and balanced realization have been introduced for the δ-operator case. In order to do this, first, the relationship between the gramians for the δ- and q-operator cases, as defined in the literature, was established. The reachability and controllability gramians, that is, P and Q, respectively, for 1-D δ-systems were found to satisfy

$$\begin{aligned} P &= \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} (cI - A_\delta)^{-1} B_\delta B_\delta^* (c^* I - A_\delta^*)^{-1} \frac{dc}{1 + \Delta c}; \\ Q &= \frac{1}{2\pi j} \oint_{\mathcal{T}_\delta} (c^* I - A_\delta^*)^{-1} C_\delta^* C_\delta (cI - A_\delta)^{-1} \frac{dc}{1 + \Delta c}, \end{aligned} \tag{T3.7}$$

where $\mathcal{T}_\delta$ is the stability boundary applicable for δ-systems, that is, $\mathcal{T}_\delta = \{c \in \Im : |c + 1/\Delta| = 1/\Delta\}$. An extension of this is then used to define the 2-D gramians of δ-systems represented in the Roesser model developed above.

For the important class of separable (that is, separable-in-denominator) systems, it is shown that these gramians may be computed through the solution of four Lyapunov equations. These notions and results are useful in many applications, such as, in extracting reduced order models of δ-systems.

Sensitivity. Measures that indicate coefficient sensitivity of the δ-models developed above have been introduced. Unlike what is available in literature, this development is applicable to the MIMO case as well. With these sensitivity measures as a guide, development of minimum sensitivity structures has been carried out. The connection with the corresponding balanced realizations has been pointed out.

Roundoff noise. With the use of a noise model that takes into account the roundoff error propagation in the s.s. model developed above, structures that minimize roundoff noise have been developed.

Publications: Work directly related to grants

[1] K. Premaratne and P.H. Bauer (1994). Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic. *Proceedings 1994 IEEE International Symposium on Circuits and Systems (ISCAS'94),* London, UK, vol. 2, 461-464.

[2] P.H. Bauer and K. Premaratne (1994). Fixed-point implementation of multi-dimensional delta-operator formulated discrete-time systems: Difficulties in convergence. *Proceedings of the 1994 IEEE SOUTHEASTCON,* Miami, FL, 26-29.

[3] K. Premaratne and A.S. Boujarwah (1994). An algorithm for stability determination of two-dimensional delta-operator formulated discrete-time systems. *Multidimensional Systems and Signal Processing,* to appear.

[4] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer (1994). Two-dimensional delta-operator formulated discrete-time systems: State-space realization and its coefficient sensitivity properties. *37th Midwest Symposium on Circuits and Systems,* Lafayette, LA, to be presented; *IEEE Transactions on Signal Processing,* in preparation.

[5] E.C. Kulasekere, K. Premaratne, P.H. Bauer, and L.J. Leclerc (1994). An exhaustive search algorithm for checking limit cycle behavior of digital filters. *IEEE Transactions on Signal Processing,* in preparation.

Note. The contents of [1] and [2] are also being prepared for possible publication in *IEEE Transactions on Signal Processing.*

Publications: Other work where grants are acknowledged

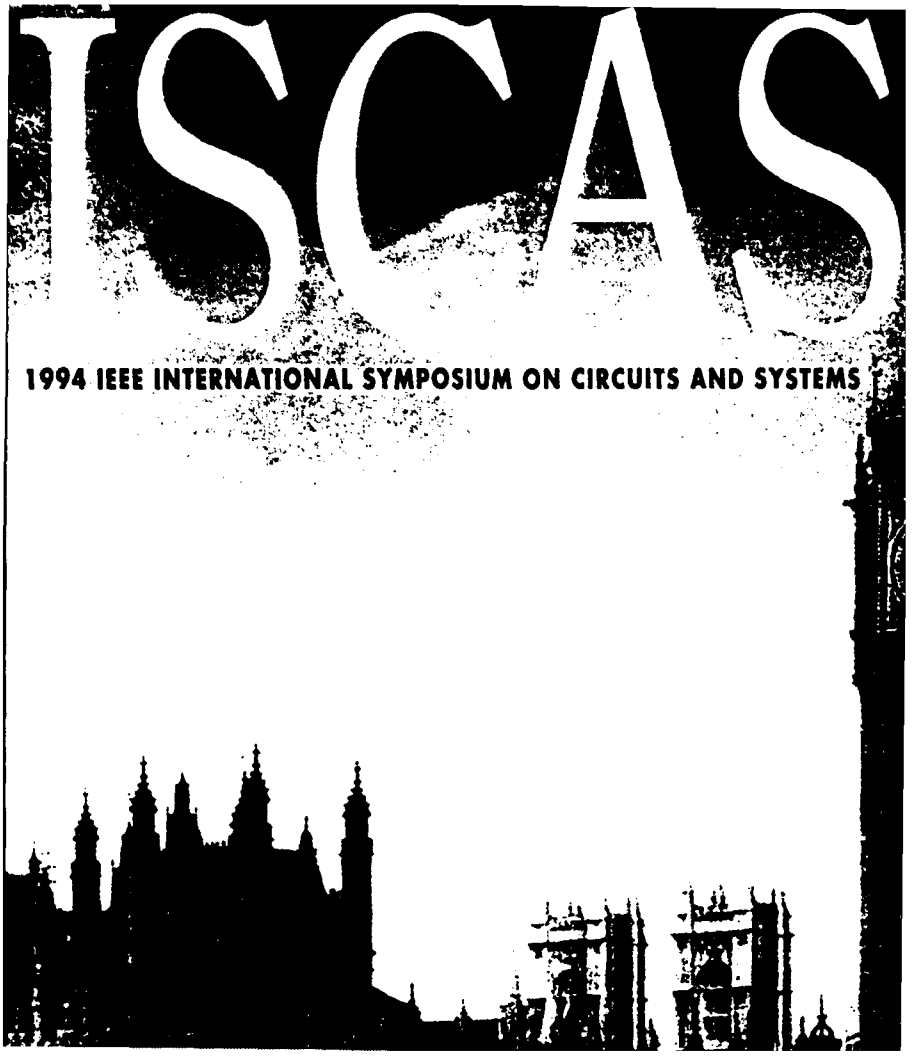
[1] K. Premaratne and E.I. Jury (1994). Discrete-time positive-real lemma revisited: The discrete-time counterpart of the Kalman-Yakubovitch lemma. *IEEE Transactions on Circuits and Systems—I. Fundamental Theory and Applications,* to appear.

[2] M.M. Ekanayake and K. Premaratne (1994). Two-channel IIR QMF filter banks with approximately linear-phase analysis and synthesis filters. *28th Annual Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, to be presented; *IEEE Transactions on Signal Processing,* in review.

[3] K. Premaratne and M. Mansour (1994). Robust stability of time-variant discrete-time systems with bounded parameter perturbations. *IEEE Transactions on Circuits and Systems—I. Fundamental Theory and Applications,* in review.

[4] S.A. Yost and P.H. Bauer (1994). Robust stability of multi-dimensional difference equations with shift-variant coefficients. *Multidimensional Systems and Signal Processing,* to appear.

Appendix: Published Papers


ISCAS
1994 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS

Limit cycles and asymptotic stability of delta-operator formulated discrete-time systems implemented in fixed-point arithmetic

Kamal Premaratne
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124
USA
(+1) 305-284-4051
kprema@umiami.ir.miami.edu

Peter H. Bauer
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556
USA
(+1) 219-631-8015
pbauer@mars.ee.nd.edu

ABSTRACT

This paper analyzes the problem of global asymptotic stability of delta-operator formulated discrete-time systems implemented in fixed-point arithmetic. It is shown that the free response of such a system tends to produce period one limit cycles if conventional quantization arithmetic schemes are used. Explicit necessary conditions for global asymptotic stability are derived, and these demonstrate that, in almost all cases, fixed-point arithmetic does not allow for global asymptotic stability in delta-operator formulated discrete-time systems that use a short sampling time.

I. INTRODUCTION

Recently, discrete-time systems formulated with the incremental difference operator (or, δ-operator) have been receiving considerable attention in the technical literature [1-4]. Most of this work focus on its superior performance under finite wordlength conditions when compared with those formulated with the shift-operator (or, q-operator). In particular, investigations of coefficient sensitivity and quantization noise properties have revealed that δ-operator formulations usually perform significantly better than their q-operator counterparts [1-4]. This is especially true for high-speed applications where the sampling rate is much larger than the underlying system bandwidth. Under these conditions, q-operator formulated discrete-time systems tend to become ill-conditioned [1-2].

Although a large amount of work is available on the effects of coefficientsensitivity and quantization noise, a deterministic study of the nonlinear behavior of discrete-time systems formulated with the δ-operator has not been undertaken. In the case of floating-point (FLP) arithmetic, some results for feedback system are available in [2].

In this work, we focus on the convergence behavior of the unforced system response and global asymptotic stability of δ-operator formulated discrete-time systems implemented in fixed-point (FXP) arithmetic. In particular, via necessary conditions for stability, it will be shown that such systems tend to produce DC limit cycles.

The structure of this article is as follows: In Section II, we introduce notation and nomenclature. The model for δ-operator formulated discrete-time systems, with and without quantization nonlinearities, is briefly discussed. Section III addresses the problem of asymptotic stability when FXP arithmetic is used for the implementation. In terms of ensuing DC limit cycles, necessary conditions for global asymptotic stability are formulated. It is shown that, when FXP arithmetic is used, stability of the linear system is often lost. Section IV provides concluding remarks.

II. NOTATION AND NOMENCLATURE

Since our focus is on investigation of stability properties of δ-operator formulated discrete-time systems under unforced conditions, the state equations of the system under zero-input will be considered.

In the linear case, the general m-th order state-space representation is given by

$$\delta[\mathbf{x}](n) = A^{\delta}\mathbf{x}(n); \tag{1}$$

$$\mathbf{x}(n+1) = \mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n), \tag{2}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_m(n)]^T$ is the state vector at instant n, $A^{\delta} = \{a_{ij}^{\delta}\} \in \Re^{m \times m}$ is the system matrix,

and $\Delta > 0$ is the sampling time. Moreover, $\delta[\cdot]$ represents the δ-operator, that is,

$$\delta[x_\nu](n) = \frac{x_\nu(n+1) - x_\nu(n)}{\Delta}, \quad \forall \nu = 1, \ldots, m, \qquad (3)$$

and $\delta[\mathbf{x}](n) = [\delta[x_1](n), \ldots, \delta[x_m](n)]^T$. The actual implementation of (1) and (2) in FXP format gives rise to nonlinear quantization operations that occur at various locations depending on the hardware realization.

Eqn. (1) can be implemented either by using single wordlength accumulators (creating a quantization error after each multiplication) or by using double wordlength accumulators (creating a quantization error only after summation). We will only consider the latter option since practically all modern DSP machines implement this. Eqn. (1) can then be written as

$$\delta[\mathbf{x}](n) = Q\{A^\delta \mathbf{x}(n)\}, \qquad (4)$$

where Q is a vector-valued quantization nonlinearity of the form

$$Q\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}. \qquad (5)$$

Here, $Q\{x_\nu\}$ denotes magnitude truncation, two's complement truncation, or rounding.

Eqn. (2) can be implemented in two different ways:

$$\mathbf{x}(n+1) = \mathbf{x}(n) + Q\{\Delta \cdot \delta[\mathbf{x}](n)\}, \qquad (6)$$

or

$$\mathbf{x} = Q\{\mathbf{x}(n) + \Delta \cdot \delta[\mathbf{x}](n)\}. \qquad (7)$$

Eqn. (6) corresponds to quantization after multiplication while (7) corresponds to quantization after summation. In contrast to (1), for (2), it is not clear which of the two quantization schemes in (6) and (7) is preferable. We will therefore consider both possibilities.

Throughout this paper, we will use the following definition of stability:

Definition. The discrete-time system in $\{(4),(6)\}$ or $\{(4),(7)\}$ is globally asymptotically stable if and only if, for any initial condition $\mathbf{x}(0)$, the state vector $\mathbf{x}$ asymptotically reaches zero, that is, $\mathbf{x}(n) \to 0$ for $n \to \infty$.

Comment. Since the FXP systems considered are in fact finite state machines, the condition $\mathbf{x}(n) \to 0$ for $n \to \infty$ may be restated as $\mathbf{x}(N) = 0$ for some finite N [5].

Finally, the symbol ℓ is used to denote the quantization step.

III. NECESSARY CONDITIONS FOR STABILITY

First, we will consider the system described by $\{(4),(6)\}$. From the definition for global asymptotic stability as stated in the previous section, it is necessary that

$$Q\{\Delta \cdot \delta[\mathbf{x}](n)\} \neq 0, \quad \text{for any} \quad \mathbf{x}(n) \neq 0. \qquad (8)$$

This is just one of a finite set of conditions that is required to ensure global asymptotic stability of a FXP implementation of a linearly stable system [5].

In the case of rounding, condition (8) is violated if

$$|\Delta \cdot \delta[x_\nu](n)| \leq \frac{\ell}{2}, \quad \text{for any} \quad \nu = 1, \ldots, m. \qquad (9)$$

The sampling time Δ in a δ-operator formulated implementation is typically very small. With $\Delta = I \cdot \ell$ and (9), we have

$$|\delta[x_\nu](n)| \leq \frac{1}{2I}, \quad \text{for any} \quad \nu = 1, \ldots, m, \qquad (10)$$

where I is a positive integer.

In the case of magnitude truncation, (10) takes the form

$$|\delta[x_\nu](n)| \leq \frac{1}{I}, \quad \text{for any} \quad \nu = 1, \ldots m. \qquad (11)$$

Accordingly, for two's complement truncation, we have

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \quad \text{for any} \quad \nu = 1, \ldots, m. \qquad (12)$$

Conditions (10-12) describe the deadband, in terms of $\delta[\mathbf{x}]$, for which a DC limit cycle occurs. Such a limit cycle can be avoided if (10-12) are satisfied by the zero vector only. In the case of rounding, we therefore require

$$\ell > \frac{1}{2I},$$

or, equivalently,

$$\Delta > \frac{1}{2}, \qquad (13)$$

which is impractical. Similarly, for magnitude and two's complement truncation, we obtain

$$\ell > \frac{1}{I} \Longleftrightarrow \Delta > 1, \qquad (14)$$

which again is equally impractical.

This result is summarized in the following theorem.

Theorem 1. A necessary condition for stability of the δ-operator formulated discrete-time system in $\{(4),(6)\}$ is $\Delta > 0.5$ for rounding and $\Delta > 1$ for truncation.

The above theorem shows that high-speed δ-operator formulated implementations that possess a small sampling time cannot be realized limit cycle free in FXP format!

A second necessary condition for the system in $\{(4),(6)\}$ can be obtained by noting that

$$\delta[\mathbf{x}](n) = 0 \tag{15}$$

can occur in (4) even though the state vector $\mathbf{x}(n) \neq 0$.

Therefore, for rounding, no nonzero state vector $\mathbf{x}(n)$ that satisfies

$$-\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \leq A^{\delta} \cdot \mathbf{x}(n) \leq +\begin{pmatrix} \frac{\ell}{2} \\ \vdots \\ \frac{\ell}{2} \end{pmatrix} \tag{16}$$

may be allowed to exist. Here, the inequality has to hold elementwise. Taking norms on both sides of (16) one gets an algebraic condition on the system matrix A^{δ} that always support DC limit cycles. Eqn. (16) has the following interesting interpretations:

1. Each of the resulting m inequalities can be geometrically interpreted as the intersection of two half spaces in $\Re^m$. These intersections are symmetric about the origin and have parallel boundaries. The normal vector to the boundaries is given by the particular row vector of A^{δ}. Only if the intersection of *all* such m half spaces contains a nonzero point in $\Re^m$, and if it belongs to the quantization lattice, will there exist a nonzero state vector that is an equilibrium point of the system.

2. Eqn. (16) can also be interpreted from an eigenvalue/eigenvector viewpoint. In high-speed digital filters where the sampling frequency is typically much higher than the bandwidth of the processed signal, a q-operator implementation's eigenvalues cluster around the point $z = 1$ [1]. The corresponding δ-operator implementation for large sampling times has eigenvalues clustered around zero. However, as the sampling time becomes small, these eigenvalues move towards the eigenvalues of the underlying continuous-time system [1]. In other words, for large sampling times, the system matrix will be ill-conditioned, that is, vectors $\mathbf{x}(n) \neq 0$ exist such that $A^{\delta} \cdot \mathbf{x}(n)$ is close to the zero vector. According to (16), this is likely to cause a DC limit cycle. For small sampling times, this problem may not occur; however, in this case, the conditions in Theorem 1 are not satisfied!

In the case of the remaining two quantization schemes, the inequalities corresponding to (16) are given as follows: For two's complement truncation,

$$0 \leq A^{\delta} \cdot \mathbf{x}(n) < \begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq 0, \tag{17}$$

and, for magnitude truncation,

$$-\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix} < A^{\delta} \cdot \mathbf{x}(n) < +\begin{pmatrix} \ell \\ \vdots \\ \ell \end{pmatrix}, \ \mathbf{x}(n) \neq 0. \tag{18}$$

A similar analysis can be conducted for the system in $\{(4),(7)\}$. Since (4) is common to both realizations, (16-18) are still valid and provide conditions under which the finite difference is quantized to zero and a DC limit cycle is produced. We will now briefly discuss necessary conditions for global asymptotic stability obtained from (7).

For rounding, proceeding as in (9), we have

$$|\Delta \cdot \delta[x_\nu](n)| \leq \frac{\ell}{2}, \quad \text{for any} \quad \nu = 1, \ldots, m,$$

and therefore

$$|\delta[x_\nu](n)| \leq \frac{1}{2I}, \quad \text{for any} \quad \nu = 1, \ldots, m. \tag{19}$$

For magnitude truncation, we obtain

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \ \forall \delta[x_\nu] \geq 0, \tag{20}$$

and

$$-\frac{1}{I} < \delta[x_\nu](n) \leq 0, \ \forall \delta[x_\nu] < 0. \tag{21}$$

In the case of two's complement truncation, the condition for a DC limit cycle is given by

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \ \forall \nu = 1, \ldots, m. \tag{22}$$

With $\Delta = I \cdot \ell$, I being a 'small' integer, we come to the same conclusion as for the previously considered system:

$$\Delta > \frac{1}{2} \quad \text{for rounding;}$$

$$\Delta > 1 \quad \text{for truncation.}$$

Therefore, Theorem 1 also holds for the system representation in $\{(4),(7)\}$.

IV. CONCLUSION

Via a set of necessary conditions for global asymptotic stability, it has been shown that high-speed, limit cycle free δ-operator implementations of linear discrete-time systems cannot be realized. This is due to the tendency of such a realization to produce period one limit cycles. This situation arises from small values in the finite difference being quantized to zero. Hence, convergence to the 'wrong' equilibrium point is very likely. Conditions on the system matrix and the sampling time if such limit cycle behavior is to be avoided have been provided. The results indicate that, in high-speed applications, these conditions cannot be satisfied with conventional quantization schemes.

ACKNOWLEDGEMENT

This work was partially supported by a research grant from the Office of Naval Research (ONR).

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High speed digital signal processing and control," *Proceedings of the IEEE*, 80, 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta-operators," *IEEE Transactions Automatic Control*, 31, 11, pp. 1015-1021, Nov. 1986.

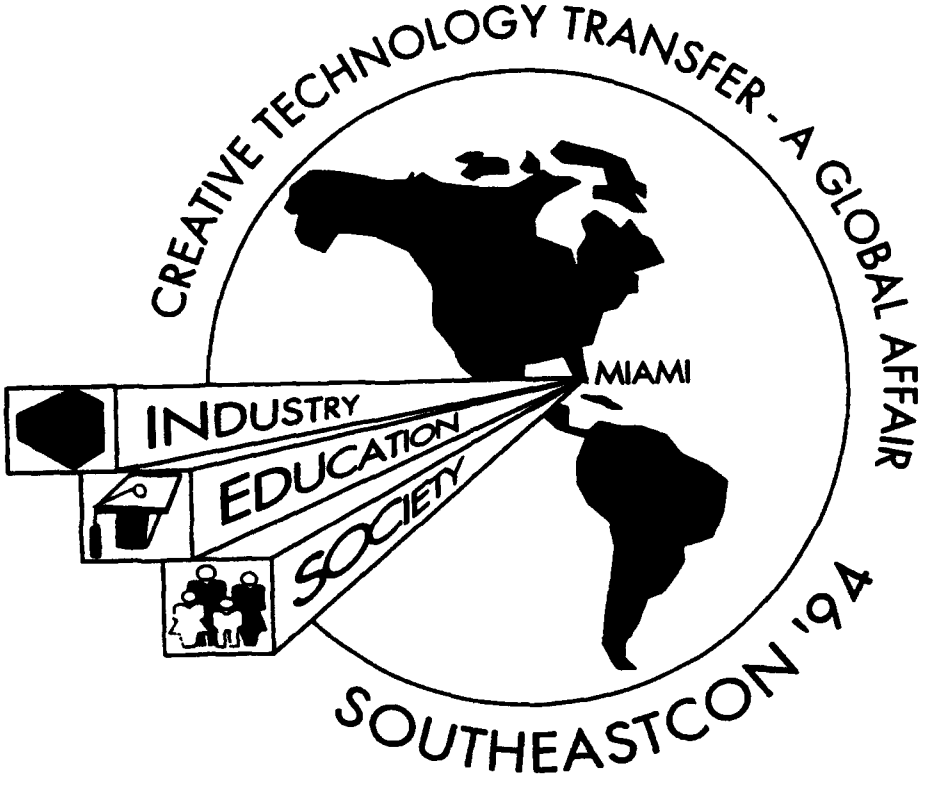
[3] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the 1990 IEEE Conference on Decision and Control (CDC'90)*, 2, pp. 954-959, Honolulu, HI, Dec. 1990.

[4] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Transactions on Signal Processing*, 41, 2, pp. 629-637, Feb. 1993.

[5] P.H. Bauer and L.J. Leclerc, "A computer-aided test for the absence of limit cycles in fixed point digital filters," *IEEE Transactions on Signal Processing*, 39, 11, pp. 2400-2410, Nov. 1991.

[6] K. Premaratne, R. Salvi, N.R. Habib, and J.P. LeGall, "Delta-operator formulated discrete-time approximations of continuous-time systems," to appear in *IEEE Transactions on Automatic Control*, 1994.

PROCEEDINGS OF 1994 IEEE SOUTHEASTCON '94



Conference and Exhibit

April 10 - 13, 1994
Miami, Florida

94CH3392-8

Hosted by:
Florida International University ECE Department
University of Miami ECE Department
IEEE Miami Section
IEEE Region 3
IEEE Florida Council

FIXED-POINT IMPLEMENTATION OF MULTI-DIMENSIONAL DELTA-OPERATOR FORMULATED DISCRETE-TIME SYSTEMS: DIFFICULTIES IN CONVERGENCE

Peter H. Bauer, PhD
Department of Electrical Engineering
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556

Kamal Premaratne, PhD
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124

Abstract— In this paper, the convergence properties of linearly stable multi-dimensional systems are investigated for the case of delta-operator implementations in fixed-point format. It is shown that zero-convergence is almost never achieved, if the sampling time is small. Using a one-dimensional analysis, it is demonstrated that zero-convergence cannot be guaranteed along the axis of the first hyper-quadrant for a first hyper-quadrant causal system. This limits the use of delta-operators for solving partial differential equations in discrete time with fixed-point arithmetic.

I. INTRODUCTION

Delta-operator (or, δ-operator) implementations of discrete-time systems have been the topic of a number of research papers within the last decade. A comprehensive treatment of the properties of δ-operator implementations can be found in [1]. It is well known that δ-operators outperform shift-operators (or, q-operators) in terms of their finite wordlength properties [2]. In particular, its quantization noise and sensitivity properties make the δ-operator an interesting alternative to the q-operator in areas such as digital control, digital signal processing, and generally discrete-time simulation of dynamical systems described by differential equations [1], [3].

In this paper, we will perform a deterministic analysis of the finite wordlength properties of multi-dimensional (m-D) δ-operator implemented discrete-time systems. In particular, we will investigate the zero-convergence of δ-operator fixed-point implementations of one-dimensional (1-D) and m-D systems. Although it is of vital importance, this problem has not been investigated thus far in the literature. After all, asymptotic stability and convergence to the true equilibrium points are some of the most fundamental requirements for any discrete-time system realization.

This article is organized in the following way: Section II introduces the notation. The m-D δ-operator model will be introduced and briefly discussed. This section will also provide the problem formulation. Section III provides necessary 1-D stability conditions for m-D first hyper-quadrant causal systems with nonlinearities. Using these necessary conditions, section IV provides a stability and convergence analysis for m-D systems. It will be shown that the resulting 1-D systems cannot ensure zero-convergence. Section V contains concluding remarks.

II. NOTATION AND PROBLEM FORMULATION

The m-D Roesser model has the following δ-operator formulation [4]:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} A^{\delta}_{11} & \cdots & A^{\delta}_{1m} \\ \vdots & \ddots & \vdots \\ A^{\delta}_{m1} & \cdots & A^{\delta}_{mm} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \begin{bmatrix} B^{\delta}_{1} \\ \vdots \\ B^{\delta}_{m} \end{bmatrix} \mathbf{u}(\mathbf{n}); \quad (1)$$

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix}. \quad (2)$$

The input-state equations in (1) and (2) describe a first hyper-quadrant causal m-D system with a uniform sampling period of Δ in all directions. The operators $q^{(i)}$ and $\delta^{(i)}$ represent the shift- and delta-operator in the direction specified by the axis n_i. In particular

$$q^{(i)}[\mathbf{x}^{(i)}](\mathbf{n}) = \mathbf{x}^{(i)}(n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_m) \quad (3a)$$

$$\delta^{(i)}[\mathbf{x}^{(i)}](\mathbf{n})$$

0-7803-1797-1/94/$3.00 © 1994 IEEE

$$= \frac{1}{\Delta}(\mathbf{x}^{(i)}(n_1,\ldots,n_{i-1},n_i+1,n_{i+1},\ldots,n_m) - \mathbf{x}^{(i)}(\mathbf{n})). \quad (3b)$$

Here, $(\mathbf{n}) \doteq (n_1,\ldots,n_m)$ denotes a point in the first hyper-quandrant, $\mathbf{x}^{(i)}(\mathbf{n})$ is the portion of the state vector propagating in the direction specified by the axis n_i, $\mathbf{u}(\mathbf{n})$ is the m-D input vector, and A_{ij}^{δ} and B_i^{δ}, for $i = 1,\ldots,m$, $j = 1,\ldots,m$, are the submatrices of the system and input matrices, respectively.

If (1) is realized in fixed-point arithmetic, it takes the following form under zero-input conditions:

$$\begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} A_{11}^{\delta} & \cdots & A_{1m}^{\delta} \\ \vdots & \ddots & \vdots \\ A_{m1}^{\delta} & \cdots & A_{mm}^{\delta} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} \right\} \quad (4)$$

where $\mathbf{Q}\{\mathbf{x}\} = \begin{pmatrix} Q\{x_1\} \\ \vdots \\ Q\{x_m\} \end{pmatrix}$ with $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$.

Equation (4) assumes quantization after summation; since practically all modern DSP machines implement this quantization scheme, we utilize this. The vector-valued quantization nonlinearity $\mathbf{Q}\{\cdot\}$ may represent any one of the conventional schemes, viz., magnitude truncation, magnitude rounding, two's complement truncation, and two's complement rounding.

Equation (2) can be implemented in two different forms:

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(i)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \mathbf{Q}\left\{ \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \quad (5)$$

or

$$\begin{bmatrix} q^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ q^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} = \mathbf{Q}\left\{ \begin{bmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta^{(1)}[\mathbf{x}^{(1)}](\mathbf{n}) \\ \vdots \\ \delta^{(m)}[\mathbf{x}^{(m)}](\mathbf{n}) \end{bmatrix} \right\} \quad (6)$$

Equation (5) corresponds to quantization after multiplication, whereas (6) corresponds to quantization after addition. In contrast to (1), for (2), it is not obvious which of the two forms stated above is preferable.

The following definition for asymptotic stability [5] will be used throughout this paper.

Definition. An m-D first hyper-quadrant causal discrete-time system is asymptotically stable under all finitely extended bounded input signals $u(\mathbf{n})$ where

$$|u(\mathbf{n})| \le S, \quad \text{for} \quad n_1 + \cdots + n_m \le D; \quad (7)$$

$$u(\mathbf{n}) = 0, \quad \text{for} \quad n_1 + \cdots + n_m > D, \quad (8)$$

if all the states of the m-D discrete-time system asymptotically reach zero for $n_1 + \cdots + n_m \to \infty$. Here, $n_\nu \ge 0$, $\nu = 1,\ldots,m$, S is a nonnegative real number, and D is a positive integer.

Since the fixed-point systems considered are in fact finite state machines, the condition

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} \to 0,$$

for $n_1 + \cdots + n_m \to \infty$, $n_\nu \ge 0$, $\nu = 1,\ldots,m$, can be strengthened to

$$\begin{pmatrix} \mathbf{x}^{(1)}(\mathbf{n}) \\ \vdots \\ \mathbf{x}^{(m)}(\mathbf{n}) \end{pmatrix} = 0,$$

for all points $n_1 + \cdots + n_m \ge c$, $n_\nu \ge 0$, $\nu = 1,\ldots,m$, where c is some finite integer.

Problem Formulation. Analyze the asymptotic zero-convergence of the state response of systems in (4,5) and (4,6) under the assumption that the underlying linear system is asymptotically stable.

III. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY OF m-D SYSTEMS

In this section, we present some necessary conditions for stability of a first hyper-quadrant causal m-D discrete-time system represented in its Roesser local state-space model in (1,2). These necessary conditions are formulated in terms of 1-D conditions. This theorem follows directly from a result in [6] which was formulated for q-operator implemented discrete-time systems. The proof of the theorem rests on the fact that a first hyper-quadrant m-D system can be described by a 1-D system for those locations that are along the m coordinate axes of the boundary of the hyper-quadrant. Reformulating the result in [6] for δ-operator systems produces the following theorem:

Theorem 1.

(a) A necessary condition for global asymptotic stability of the system in (4,5) is that each of the following 1-D systems in (9,10) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \quad (9)$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{x}^{(i)}(n_i) + \mathbf{Q}\left\{\Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\} \quad (10)$$

where $i = 1, \ldots, m$.

(b) A necessary condition for global asymptotic stability of the system in (4,6) is that each of the following in 1-D systems in (11,12) is globally asymptotically stable:

$$\delta^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{[A_{ii}^{\delta}]\mathbf{x}^{(i)}(n_i)\right\}; \quad (11)$$

$$q^{(i)}[\mathbf{x}^{(i)}](n_i) = \mathbf{Q}\left\{\mathbf{x}^{(i)}(n_i) + \Delta \cdot \delta^{(i)}[\mathbf{x}^{(i)}](n_i)\right\} \quad (12)$$

where $i = 1, \ldots, m$.

Proof. **For a detailed proof, and generalizations to higher sub-dimensional systems, the reader is referred to [6].** ■

Theorem 1 can be viewed as an extension of the concept of practical BIBO stability to asymptotic stability of nonlinear systems. It is particularly useful in proving instability in m-D nonlinear systems.

IV. NECESSARY CONDITIONS FOR GLOBAL ASYMPTOTIC STABILITY OF 1-D SYSTEMS

Let us rewrite (9), (10), and (12) as 1-D matrix equations of order K. In this case, (9), (10), and (12) yield (13), (14), and (15), respectively:

$$\begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix} = \mathbf{Q}\left\{\begin{bmatrix} a_{11}^{\delta} & \cdots & a_{1K}^{\delta} \\ \vdots & \ddots & \vdots \\ a_{n1}^{\delta} & \cdots & a_{KK}^{\delta} \end{bmatrix}\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix}\right\}; \quad (13)$$

$$\begin{bmatrix} x_1(n+1) \\ \vdots \\ x_K(n+1) \end{bmatrix} = \begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \mathbf{Q}\left\{\Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix}\right\}; \quad (14)$$

$$\begin{bmatrix} x_1(n+1) \\ \vdots \\ x_K(n+1) \end{bmatrix} = \mathbf{Q}\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix}\right\}. \quad (15)$$

Now, we are in a position to formulate the second theorem which presents a necessary condition for stability of 1-D systems.

Theorem 2. **A necessary condition for global asymptotic stability of the system in (13,14) or (13,15) is given by**

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$
$$\Delta \geq 1, \quad \text{for truncating.}$$

Proof. **For global asymptotic stability of (13,14), it is necessary that**

$$\mathbf{Q}\left\{\Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix}\right\} \neq 0, \quad (16)$$

for any $\begin{pmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{pmatrix} \neq 0.$

First, we will address the case of magnitude rounding. Obviously, condition (16) is violated if, for $x_\nu \neq 0$,

$$|\Delta \cdot \delta[x_\nu](n)| < \frac{\ell}{2}, \quad \text{for} \quad \nu = 1, \ldots, K, \quad (17)$$

where ℓ is the quantization step. Expressing the sampling time Δ as an integer multiple of ℓ, we have

$$\Delta = I \cdot \ell, \quad (18)$$

where I is some (typically small) positive integer. With (17) and (18), we obtain the following condition for instability:

$$|\delta[x_\nu](n)| < \frac{1}{2I}, \ \nu = 1, \ldots, m, \quad (19)$$

for $x_\nu \neq 0$, $\nu = 1, \ldots, m$.

Condition (19) is not satisfied for any nonzero value of x_ν (that is, the condition for instability is not satisfied) if $\ell \geq 1/2I$, or equivalently,

$$\Delta \geq \frac{1}{2}. \quad (20)$$

This proves the theorem for magnitude rounding.

For the case of magnitude truncating, (17) takes the form

$$|\Delta \cdot \delta[x_\nu](n)| < \ell, \quad \text{for} \quad \nu = 1, \ldots, K. \quad (21)$$

Therefore, (19) becomes

$$|\delta[x_\nu](n)| < \frac{1}{I}. \quad (22)$$

This finally yields

$$\Delta \geq 1. \tag{23}$$

For two's complement, (17) takes the form

$$0 \leq \Delta \cdot \delta[x_\nu](n) < \ell, \quad \text{for} \quad \nu = 1, \ldots, K. \tag{24}$$

This results in

$$0 \leq \delta[x_\nu](n) < \frac{1}{I}, \tag{25}$$

and consequently, $\Delta \geq 1$. This proves the theorem for the system in (13,14). A similar argument can be used for the system in (13,15) by considering the cases for which

$$Q\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix} + \Delta \cdot \begin{bmatrix} \delta[x_1](n) \\ \vdots \\ \delta[x_K](n) \end{bmatrix}\right\} = Q\left\{\begin{bmatrix} x_1(n) \\ \vdots \\ x_K(n) \end{bmatrix}\right\}, \tag{26}$$

for nonzero state vectors. ■

We can now combine Theorems 1 and 2 to formulate a necessary condition for stability of m-D first hyperquadrant causal δ-operator formulations of the generalized Roesser model.

Corollary 3. A necessary condition for global asymptotic stability of the m-D systems in (4,5) or (4,6) is

$$\Delta \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta \geq 1, \quad \text{for truncatiing.}$$

Proof. The proof follows from Theorems 1 and 2. ■

Comments.

1. Theorem 2 and Corollary 3 are also essentially applicable to the case where the sampling time varies with the direction of propagation. In this case, the inequalities in Theorem 2 and Corollary 3 would have to be replaced by

$$\Delta_i \geq 0.5, \quad \text{for magnitude rounding;}$$

$$\Delta_i \geq 1, \quad \text{for truncating,}$$

for $i = 1, \ldots, m$.

2. Most of the previous results on the superior finite wordlength properties of δ-operators depend on choosing a very small sampling time Δ. In such a case, Theorem 2 and Corollary 3 show that the system response will not converge to zero for the unforced case.
3. Our analysis is limited to the zero-input case for which DC limit cycles were used to derive conditions for non-convergence. If one includes other types of limit cycles in the analysis, the requirements for Δ may become even more severe.
4. Theorem 2 and Corollary 3 show that fixed-point implementations of 1-D and m-D δ-operator systems *cannot be realized limit cycle free, if good coefficient sensitivity and quantization noise measures have to be achieved.* See also [7].

V. CONCLUSION

In this paper, it was shown that fixed-point implementations of 1-D and m-D δ-operator systems are not limit cycle free even if the underlying linear system is stable and the sampling time is chosen small. This nonconvergent behavior can be explained by the quantization of the δ-term to zero which leaves the state vector unchanged. The smaller the sampling time, the more severe this effect is. Therefore, the practical value of δ-operators for fixed-point implementations of 1-D and m-D systems is questionable. There are however indications that this effect is much less severe in floating-point implementations.

δ-operator implemented discrete-time systems represent a class of systems where the quantization noise at the output can be small compared to other realizations. However, as was shown above, such realizations will invariably exhibit limit cycle, that is, highly correlated quantization noise, behavior. Therefore, in this case, typical measures for quantization noise are of very limited use for obtaining any insight into the likelihood of limit cycles and vice versa.

ACKNOWLEDGEMENT

This work was partially supported by a grant from the Office of Naval Research (ONR).

REFERENCES

[1] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proceedings of the IEEE*, vol. 80, no. 2, pp. 240-259, Feb. 1992.

[2] R.H. Middleton and G.C. Goodwin, "Improved finite wordlength characteristics in digital control using delta operators," *IEEE Transactions on Automatic Control*, vol. 31, pp. 1015-1021, Nov. 1986.

[3] G.Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterization," *Proceedings of the IEEE Conference on Decision and Control (CDC'90)*, vol. 2, pp. 954-959, Honolulu, HI, 1990.

[4] K. Premaratne, J. Suarez, M.M. Ekanayake, and P.H. Bauer, "Delta-operator formulated implementation of two-dimensional discrete-time systems," in preparation.

[5] P. Bauer, "Finite wordlength effects in m-D digital filters with singularities on the stability boundary," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 894-900, Apr. 1992.

[6] P. Bauer, "A set of necessary stability conditions for m-D nonlinear digital filters," to appear in *Circuits, Systems and Signal Processing*, 1994.

[7] K. Premaratne and P.H. Bauer, "Limit cycles and asymptotic stability of delta-operator systems in fixed-point arithmetic," submitted to be presented at the *1994 IEEE Symposium on Circuits and Systems (ISCAS'94)*, London, UK, 1994.

Two-Dimensional Delta-Operator Formulated Discrete-Time Systems: State-Space Realization and Its Coefficient Sensitivity Properties

K. Premaratne, J. Suarez, and M.M. Ekanayake
Department of E&CE
University of Miami
Coral Gables, FL 33124 USA

P.H. Bauer
Department of EE
Laboratory of Image and Signal Analysis
University of Notre Dame
Notre Dame, IN 46556 USA

Abstract—By developing the δ-operator analog of the Roesser model, state-space realization of two- and multi-dimensional δ-systems is investigated. The corresponding notions of gramians and balanced realization are also defined. It is shown that, discrete-time system implementation using this model can yield superior coefficient sensitivity properties.

I. Introduction

Judging by its performance in the one-dimensional (1-D) case [2], [5-6], one is led to expect superior coefficient sensitivity and roundoff noise performance with δ-operator implementation of two-dimensional (2-D) and multi-dimensional (m-D) discrete-time (DT) systems. With this in mind, δ-operator analog of the q-operator Roesser local state-space (s.s.) model [12] is developed. We also propose the notions of gramians and balanced (BL) realization. As expected, realization using this model can provide superior coefficient sensitivity properties.

II. Nomenclature and Preliminaries

A. Nomenclature

$\Re$: Reals; $\Im$: Complex numbers; $\Re^{q\times p}$, $\Im^{q\times p}$: Matrices of size $q \times p$ over $\Re$ and $\Im$; I_n: $n \times n$ unit matrix; A^*, trace$[A]$, $\|A\|_F$: Conjugate transpose, trace, and Fröbenius norm of matrix A; $\mathbf{e}_i^{(n)}$: Unit vector in $\Re^n$ with 1 on the i-th row; $E_{i,j}^{q\times p} = \mathbf{e}_i^{(q)}\mathbf{e}_j^{(p)^*} \in \Re^{q\times p}$; $\overline{U}_{q\times p} = \sum_{i=1}^{q}\sum_{j=1}^{p} E_{i,j}^{(q\times p)} \otimes E_{i,j}^{(q\times p)} \in \Re^{q^2\times p^2}$.

For q- and δ-systems, we use the indeterminates z and c, respectively. For 1-D systems, $\delta = (q-1)/\tau \Longleftrightarrow c = (z-1)/\tau$, where τ is a positive real constant, usually the sampling time. Let $\overline{U}_\delta^2 = \{(c_h, c_v) \in \Im^2 : |c_h + 1/\tau_h| \le 1/\tau_h, |c_v + 1/\tau_v| \le 1\}$. $\mathcal{T}_\delta^2$ is its boundary. The corresponding q-system regions are denoted with the subscript q.

K.P. and P.H.B. gratefully acknowledge the support received from the Office of Naval Research (ONR) through the grants N00014-94-1-0454 and N00014-94-1-0387, respectively.

B. Preliminaries

Consider a linear, shift-invariant, strictly causal, p-input q-output 2-D DT system. Its n_hh-n_vv Roesser local s.s. model $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ takes the form [12]:

$$\begin{aligned}\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j)\end{bmatrix} &= [\hat{A}]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [\hat{B}]\mathbf{u}(i,j);\\ \mathbf{y}(i,j) &\doteq [\hat{C}]\begin{bmatrix}\mathbf{x}^h(i,j)\\ \mathbf{x}^v(i,j)\end{bmatrix} + [\hat{D}]\mathbf{u}(i,j),\end{aligned} \tag{2.1}$$

where $\mathbf{u} \in \Re^p$, $\mathbf{x}^h \in \Re^{n_h}$, $\mathbf{x}^v \in \Re^{n_v}$, and $\mathbf{y} \in \Re^q$. $\mathbf{x}^h$ and $\mathbf{x}^v$ are the h.p. and v.p. local state vectors. Take $n = n_h + n_v$. Also,

$$q_h[\mathbf{x}](i,j) = \mathbf{x}(i+1,j); \quad q_v[\mathbf{x}](i,j) = \mathbf{x}(i,j+1). \tag{2.2}$$

In what follows, we use matrix partitioning that conform to $A \doteq \begin{bmatrix}\hat{A}^{(1)} & \hat{A}^{(2)}\\ \hat{A}^{(3)} & \hat{A}^{(4)}\end{bmatrix}$, $B \doteq \begin{bmatrix}\hat{B}^{(1)}\\ \hat{B}^{(2)}\end{bmatrix}$, and $C \doteq \left[\hat{C}^{(1)}, \hat{C}^{(2)}\right]$. The corresponding 2-D characteristic equation and transfer function are

$$\begin{aligned}\det[I_z - \hat{A}] &= \det[z_h I_{n_h} \oplus z_v I_{n_v} - \hat{A}];\\ \hat{H}(z_h, z_v) &= \hat{C}(I_z - \hat{A})^{-1}\hat{B} + \hat{D},\end{aligned} \tag{2.3}$$

where $z_h, z_v \in \Im$ $I_z \doteq z_h I_{n_h} \oplus z_v I_{n_v} \in \Im^{n\times n}$. With no nonessential singularities of the second kind (NSSK) on $\mathcal{T}_q^2$, $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ is BIBO stable iff [3]

$$\det[I_z - \hat{A}] \neq 0, \ \forall (z_h, z_v) \in \overline{U}_q^2. \tag{2.4}$$

III. 2-D δ-Model

A. Local s.s. model

Analogous to the 1-D case, define $\delta_h[\cdot]$ and $\delta_v[\cdot]$ as

$$\begin{aligned}\delta_h[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i+1,j) - \mathbf{x}(i,j)}{\tau_h} = \frac{q_h[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\tau_h};\\ \delta_v[\mathbf{x}](i,j) &= \frac{\mathbf{x}(i,j+1) - \mathbf{x}(i,j)}{\tau_v} = \frac{q_v[\mathbf{x}](i,j) - \mathbf{x}(i,j)}{\tau_v}.\end{aligned} \tag{3.1}$$

Here τ_h and τ_v are positive real constants denoting the 'sampling times' along h.p. and v.p. directions, respectively. Note that

$$q_h = 1 + \tau_h \delta_h; \quad q_v = 1 + \tau_v \delta_v, \tag{3.2}$$

and letting $\tau = [\tau_h I_{n_h} \oplus \tau_v I_{n_v}] \in \Re^{n \times n}$,

$$\begin{bmatrix} q_h[\mathbf{x}^h](i,j) \\ q_v[\mathbf{x}^v](i,j) \end{bmatrix} = I_n + \tau \begin{bmatrix} \delta_h I_{n_h} & 0 \\ 0 & \delta_v I_{n_v} \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}. \tag{3.3}$$

Using (3.3) in (2.1), we get

$$\begin{aligned} \begin{bmatrix} \delta_h[\mathbf{x}^h](i,j) \\ \delta_v[\mathbf{x}^v](i,j) \end{bmatrix} &\doteq [A] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [B]\mathbf{u}(i,j); \\ \mathbf{y}(i,j) &\doteq [C] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} + [D]\mathbf{u}(i,j), \end{aligned} \tag{3.4}$$

where $A \doteq \begin{bmatrix} A^{(1)} & A^{(2)} \\ A^{(3)} & A^{(4)} \end{bmatrix}$, $B \doteq \begin{bmatrix} B^{(1)} \\ B^{(2)} \end{bmatrix}$, and $C \doteq \left[C^{(1)}, C^{(2)}\right]$. In addition, we need to perform

$$q_h[\mathbf{x}^h] = \mathbf{x}^h + \tau_h \cdot \delta_h[\mathbf{x}^h]; \; q_v[\mathbf{x}^v] = \mathbf{x}^v + \tau_v \cdot \delta_v[\mathbf{x}^v]. \tag{3.5}$$

Here,

$$\hat{A} = I_n + \tau A; \; \hat{B} = \tau B; \; \hat{C} = C; \; \hat{D} = D. \tag{3.6}$$

B. Properties of the 2-D δ-model

Most of the following properties may be derived in a manner that is exactly analogous to that in [12].

The *transition matrix* $A^{i,j}$ of the δ-model, may be recursively computed from

$$A^{i,j} = \begin{cases} 0, (i,j) = (0,0); \\ [I_{n_h} \oplus I_{n_v}], (i,j) = (0,0); \\ \begin{bmatrix} I_{n_h} & 0 \\ 0 & 0 \end{bmatrix} + \tau \begin{bmatrix} A^{(1)} & A^{(2)} \\ 0 & 0 \end{bmatrix}, (i,j) = (1,0); \\ \begin{bmatrix} 0 & 0 \\ 0 & I_{n_v} \end{bmatrix} + \tau \begin{bmatrix} 0 & 0 \\ A^{(3)} & A^{(4)} \end{bmatrix}, (i,j) = (0,1); \\ A^{1,0}A^{i-1,j} + A^{0,1}A^{i,j-1}, \text{ elsewhere.} \end{cases} \tag{3.7}$$

The general response of the δ-model is

$$\begin{aligned} \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix} &= \sum_{k=0}^{j} A^{i,j-k} \begin{bmatrix} \mathbf{x}^h(0,k) \\ 0 \end{bmatrix} \\ &\quad + \sum_{h=0}^{i} A^{i-j,j} \begin{bmatrix} 0 \\ \mathbf{x}^v(h,0) \end{bmatrix} + \mathbf{f}(\mathbf{u}), \end{aligned} \tag{3.8}$$

where $\mathbf{f}(\mathbf{u}) = \sum_{(0,0) \le (h,k) < (i,j)} (A^{i-h-1,j-k}\tau \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i-h,j-k-1}\tau \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix})\mathbf{u}(h,k)$.

Let $I_c \doteq c_h I_{n_h} \oplus c_v I_{n_v} \in \Im^{n \times n}$. Then, the 2-D δ-model's characteristic equation and transfer function are

$$\begin{aligned} \det[I_c - A] &= \frac{1}{\det[\tau]} \det[I_z - \hat{A}]|_{z \to c}; \\ H(c_h, c_v) &= \hat{H}(z_h, z_v)|_{z \to c}, \end{aligned} \tag{3.9}$$

where

$$z_h = 1 + \tau_h c_h; \quad z_v = 1 + \tau_v c_v. \tag{3.10}$$

From now on, the variable transformation in (3.10) is denoted by $\mathbf{c} \to \mathbf{z}$ or $\mathbf{z} \to \mathbf{c}$ whatever is appropriate.

Nonsingular transformations of the type

$$\begin{bmatrix} \tilde{\mathbf{x}}^h(i,j) \\ \tilde{\mathbf{x}}^v(i,j) \end{bmatrix} = [T] \begin{bmatrix} \mathbf{x}^h(i,j) \\ \mathbf{x}^v(i,j) \end{bmatrix}, \tag{3.11}$$

where $T \doteq [T^{(1)} \oplus T^{(4)}]$, yield the equivalent 2-D s.s. realization $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$. Here,

$$\tilde{A} = TAT^{-1}; \; \tilde{B} = TB; \; \tilde{C} = CT^{-1}; \; \tilde{D} = D. \tag{3.12}$$

With no NSSK on $\mathcal{T}_\delta^2$, $\{A, B, C, D\}$ is BIBO stable iff

$$\det[I_c - A] \neq 0, \; \forall (c_h, c_v) \in \overline{U}_\delta^2. \tag{3.13}$$

C. Gramians

The gramians of 2-D q-systems are taken to be natural extensions of the integral expressions of their 1-D counterparts [11]. We will adopt a similar approach. In what follows, we consider the 1-D (or 2-D) stable δ-system $\{A, B, C, D\}$ with gramians P and Q. The corresponding q-system is $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ with gramians $\hat{P}$ and $\hat{Q}$.

1-D case. The gramians are defined in [10].

Definition 3.1. [10]. The gramians are the solutions to the Lyapunov equations

$$\begin{aligned} AP + PA^* + \tau \cdot APA^* &= -BB^*; \\ A^*Q + QA + \tau \cdot A^*QA &= -C^*C. \end{aligned}$$

Lemma 3.1. The gramians satisfy the integral expressions

$$P = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} FF^* \frac{dc}{1 + \tau c}; \; Q = \frac{1}{2\pi j} \oint_{\mathcal{T}_q} G^*G \frac{dc}{1 + \tau c},$$

where $F(c) \doteq (cI_n - A)^{-1}B$ and $G(c) \doteq C(cI_n - A)^{-1}$. Moreover, $\hat{P} = \tau P$ and $\hat{Q} = Q/\tau$.

Proof. Substitute $\hat{A} = I_n + \tau A$, $\hat{B} = \tau B$, $\hat{C} = C$, and $\hat{D} = D$ [10] in the equations in Definition 3.1, and note the integral expressions for P and Q in [8]. ■

2-D case. With Lemma 3.1 in mind, we have

Definition 3.2. The gramians are

$$P = \frac{1}{(2\pi j)^2} \oint_{T_\delta^2} FF^* \frac{dc_h}{1+\tau_h c_h} \frac{dc_v}{1+\tau_v c_v};$$

$$Q = \frac{1}{(2\pi j)^2} \oint_{T_\delta^2} G^*G \frac{dc_h}{1+\tau_h c_h} \frac{dc_v}{1+\tau_v c_v},$$

where $P \doteq \begin{bmatrix} P^{(1)} & P^{(2)} \\ P^{(3)} & P^{(4)} \end{bmatrix}$ and $Q \doteq \begin{bmatrix} Q^{(1)} & Q^{(2)} \\ Q^{(3)} & Q^{(4)} \end{bmatrix}$. Also, $F(c_h, c_v) \doteq (I_c - A)^{-1}B = [\mathbf{f}_1, \ldots, \mathbf{f}_n]^*$ and $G(c_h, c_v) \doteq C(I_c - A)^{-1} = [\mathbf{g}_1, \ldots, \mathbf{g}_n]$.

Remarks.

1. Note that, $(I_c - A)^{-1}|_{c \to z} = (I_z - \hat{A})^{-1}\tau$, and

$$F|_{c \to z} = \hat{F};\ G|_{c \to z} = \hat{G} \cdot \tau. \tag{3.14}$$

2. Definition 3.2 is completely analogous to the 1-D and 2-D q-systems [7], [11].

Lemma 3.2. $\hat{P} = \tau_h \tau_v P$ and $\hat{Q} = \tau_h \tau_v \tau^{-1} Q \tau^{-1}$.

Proof. Consider P in Definition 3.2. Use $c \to z$, (3.14), and definition of gramians for 2-D q-systems [11]. ■

The following are in complete analogy with 2-D q-systems.

Lemma 3.3. The gramians may be represented as

$$P = \frac{1}{\tau_h \tau_v} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} M_{i,j} M_{ij}^*;$$

$$Q = \frac{1}{\tau_h \tau_v} \tau \cdot \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A^{i,j^*} C^* C A^{i,j} \cdot \tau,$$

where, for $(i,j) = (0,0)$, $M_{i,j} = 0$, and, for $(i,j) > (0,0)$, $M_{ij} = A^{i-1,j}\tau \begin{bmatrix} B^{(1)} \\ 0 \end{bmatrix} + A^{i,j-1}\tau \begin{bmatrix} 0 \\ B^{(2)} \end{bmatrix}$.

Lemma 3.4. Let $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ with gramians $\tilde{P}$ and $\tilde{Q}$ be an equivalent system as in (3.10-11). Then, $\tilde{P} = TPT^*$ and $\tilde{Q} = T^{-1^*} Q T^{-1}$. Moreover, the eigenvalues of PQ and $\tilde{P}\tilde{Q}$ are invariant.

Definition 3.3. $\{A, B, C, D\}$ is said to be *balanced* if $P^{(1)} = Q^{(1)} \doteq \Sigma^{(1)} = \text{diag}\{\sigma_1^{(1)}, \sigma_2^{(1)}, \ldots, \sigma_{n_h}^{(1)}\}$ and $P^{(4)} = Q^{(4)} \doteq \Sigma^{(4)} = \text{diag}\{\sigma_1^{(4)}, \sigma_2^{(4)}, \ldots, \sigma_{n_v}^{(4)}\}$.

If the diagonal submatrices of P and Q are each positive definite (p.d.), a BL realization may be obtained [4]. Regarding this, we have

Lemma 3.5. Local reachability and observability of $\{A, B, C, D\}$ and $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}\}$ are equivalent. Moreover, when $\{A, B, C, D\}$ is locally reachable and observable, $P^{(1)}$, $P^{(4)}$, $Q^{(1)}$, and $Q^{(4)}$ are each p.d.

Separable systems. A separable (in denominator) 2-D q-system will have $\hat{A}^{(2)} = 0$ (and/or $\hat{A}^{(3)} = 0$) and all off-diagonal submatrices of $\hat{P}$ and $\hat{Q}$ are zero. The diagonal submatrices may be computed through two pairs of Lyapunov equations [11]. Clearly, a separable 2-D q-system yields a separable 2-D δ-system.

Theorem 3.6. Let $\{A, B, C, D\}$ be separable with $A^{(2)} = 0$. Then, $P^{(2)} = Q^{(2)} = 0$ and $P^{(3)} = Q^{(3)} = 0$, and

$$\begin{aligned}
&A^{(1)}P^{(1)} + P^{(1)}A^{(1)^*} + \tau_h A^{(1)} P^{(1)} A^{(1)^*} \\
&\quad = -B^{(1)}B^{(1)^*}/\tau_v; \\
&A^{(1)^*}Q^{(1)} + Q^{(1)}A^{(1)} + \tau_h A^{(1)^*} Q^{(1)} A^{(1)} \\
&\quad = -\begin{bmatrix} C^{(1)} & R^{(4)}A^{(3)} \end{bmatrix}^* \begin{bmatrix} C^{(1)} & R^{(4)}A^{(3)} \end{bmatrix} / \tau_v; \\
&A^{(4)}P^{(4)} + P^{(4)}A^{(4)^*} + \tau_v A^{(4)} P^{(4)} A^{(4)^*} \\
&\quad = -\begin{bmatrix} B^{(2)} & A^{(3)}S^{(1)} \end{bmatrix} \begin{bmatrix} B^{(2)} & A^{(3)}S^{(1)} \end{bmatrix}^* / \tau_h; \\
&A^{(4)^*}Q^{(4)} + Q^{(4)}A^{(4)} + \tau_v A^{(4)^*} Q^{(4)} A^{(4)} \\
&\quad = -C^{(2)^*} C^{(2)} / \tau_h.
\end{aligned}$$

Here, $R^{(4)^*} R^{(4)} \doteq \tau_h \tau_v Q^{(4)}$ and $S^{(1)} S^{(1)^*} \doteq \tau_h \tau_v P^{(1)}$.

IV. Coefficient Sensitivity

By generalizing a certain sensitivity measure, Lutz and Hakimi [9] have addressed sensitivity minimization of MIMO 1-D CT systems. The SISO 2-D q-operator case is in [7]. In what follows, we study the coefficient sensitivity of the 2-D δ-model in section III. We follow a more direct approach using Kronecker product formulation and, hence, the results are applicable to the more general MIMO case. Using [1], we may show

$$S_A(c_h, c_v) = [I_n \otimes G] \cdot \overline{U}_{n \times n} \cdot [I_n \otimes F] \tag{4.1}$$

$$S_B(c_h, c_v) = [I_n \otimes G] \cdot \overline{U}_{n \times p} \tag{4.2}$$

$$S_C(c_h, c_v) = \overline{U}_{q \times n} \cdot [I_n \otimes F] \tag{4.3}$$

$$S_D(c_h, c_v) = \overline{U}_{q \times p} \tag{4.4}$$

Lemma 4.1. The quantities in (4.1-4.4) are given as

$$S_A = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} [\mathbf{f}_1^* \quad \cdots \quad \mathbf{f}_n^*];$$

$$S_B = \begin{bmatrix} \mathbf{g}_1^{(1)} & \cdots & \mathbf{g}_1^{(p)} \\ \vdots & \ddots & \vdots \\ \mathbf{g}_n^{(1)} & \cdots & \mathbf{g}_n^{(p)} \end{bmatrix};$$

$$S_C = \begin{bmatrix} \mathbf{f}_1^{(1)^*} & \cdots & \mathbf{f}_n^{(1)^*} \\ \vdots & \ddots & \vdots \\ \mathbf{f}_1^{(q)^*} & \cdots & \mathbf{f}_n^{(q)^*} \end{bmatrix};$$

$$S_D = \begin{bmatrix} E_{1,1} & \cdots & E_{1,p} \\ \vdots & \ddots & \vdots \\ E_{q,1} & \cdots & E_{q,p} \end{bmatrix}.$$

Here, $\mathbf{f}_i^{(j)^*}$ denotes a $(q \times p)$ null matrix except its j-th row which is $\mathbf{f}_i^*$ and $\mathbf{g}_i^{(j)}$ denotes a $(q \times p)$ null matrix except its j-th column which is $\mathbf{g}_i$.

Proof. This may be shown through the results in [1] and simple yet tedious algebraic manipulations. ■

Corollary 4.2. The quantities S_A, S_B, S_C, and S_D of the δ-model and the quantities $\hat{S}_{\hat{A}}, \hat{S}_{\hat{B}}, \hat{S}_{\hat{C}}$, and $\hat{S}_{\hat{D}}$ of the corresponding q-model are related by $S_A|_{\mathbf{c}\to\mathbf{s}} = \tau\hat{S}_{\hat{A}}$, $S_B|_{\mathbf{c}\to\mathbf{s}} = \tau\hat{S}_{\hat{B}}$, $S_C|_{\mathbf{c}\to\mathbf{s}} = \hat{S}_{\hat{C}}$, and $S_D|_{\mathbf{c}\to\mathbf{s}} = \hat{S}_{\hat{D}}$, where $\tau = \tau_h I_{n_h q} \oplus \tau_v I_{n_v q} \in \Re^{nq\times nq}$.

Proof. Apply (3.14) to Lemma 4.1. ■

To proceed further, we utilize the following

Definition 4.1. Let $H(\mathbf{c}_h, \mathbf{c}_v)$ be a bivariate matrix-valued function that is analytic on $\mathcal{T}_q^2$. Then,

$$\|H(\mathbf{c}_h, \mathbf{c}_v)\|_p^p \doteq \frac{1}{(2\pi j)^2} \oint_{\mathcal{T}_q^2} \|H(\mathbf{c}_h, \mathbf{c}_v)|_{\mathbf{c}\to\mathbf{s}}\|_F^p \frac{dz_h}{z_h}\frac{dz_v}{z_v}.$$

Remark. This norm is extensively utilized in related work [7] due mainly to the fact that it leads to tractable results. This, and our desire to make a comparison with the corresponding q-model, are the primary reasons for its use here.

We now define the absolute sensitivity measure

$$M \doteq \|S_A\|_1^2 + \frac{1}{p}\|S_B\|_2^2 + \frac{1}{q}\|S_C\|_2^2 + \frac{1}{pq}\|S_D\|_2^2. \quad (4.5)$$

Remarks.

1. The use of different norms is for mathematical feasibility and tractability [7], [5].
2. The weights associated with each term in (4.5) may be thought of as *averaging factors per input/output.*
3. Due to (3.5), M should contain $\|S_{\tau_h}\|$ and $\|S_{\tau_v}\|$. However, we assume that τ_h and τ_v are selected such that each possess exact binary representations. Hence, these additional terms are neglected.

Using an argument similar to that in [7], one may show the following:

$$\|S_A\|_1^2 \le \text{trace}[\hat{P}] \cdot \text{trace}[\tau\hat{Q}\tau] \quad (4.6)$$

$$\|S_B\|_2^2 = p \cdot \text{trace}[\tau\hat{Q}\tau] \quad (4.7)$$

$$\|S_C\|_2^2 = q \cdot \text{trace}[\hat{P}] \quad (4.8)$$

$$\|S_D\|_2^2 = pq \quad (4.9)$$

Combining (4.5) with (4.6-9), we get

$$M \le \overline{M} \doteq (\text{trace}[\hat{P}] + 1)(\text{trace}[\tau\hat{Q}\tau] + 1). \quad (4.10)$$

It is customary to perform a minimization of $\overline{M}$. Hence, one attempts to characterize those $\{\check{A}, \check{B}, \check{C}, \check{D}\}$ that are 'bound optimal' with respect to M. Analogous to 2-D q-systems case [7], one may for instance show that a BL realization (modulo an orthogonal nonsingular transformation) is 'bound optimal' with respect to M.

Compared to a q-system, its δ-system counterpart yields a smaller $\overline{M}$ whenever $\text{trace}[\hat{Q}] > \text{trace}[\tau\hat{Q}\tau]$, that is,

$$(1 - \tau_h^2) \cdot \text{trace}[\hat{Q}^{(1)}] + (1 - \tau_v^2) \cdot \text{trace}[\hat{Q}^{(4)}] > 0. \quad (4.11)$$

Note that, with the local reachability and observability assumption of $\{A, B, C, D\}$, p.d. of $Q^{(1)}$ and $Q^{(4)}$ (and hence of $\hat{Q}^{(1)}$ and $\hat{Q}^{(4)}$) are guaranteed. Thus, (4.11) is satisfied if $\tau_h < 1$ and $\tau_v < 1$.

VII. Conclusion

We have developed the δ-operator analog of the Roesser local s.s. model. Notions of gramians and BL realization are also proposed. As is expected, under mild conditions, this model offers superior coefficient sensitivity properties.

References

[1] J.W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circ. Syst.*, vol. CAS-25, pp. 772-781, Sept. 1978.

[2] G.C. Goodwin, R.H. Middleton, and H.V. Poor, "High-speed digital signal processing and control," *Proc. IEEE*, vol. 80, pp. 240-259, 1992.

[3] E.I. Jury, "Stability of multidimensional systems and other related problems," in *Multidimensional Systems, Techniques, and Applications*, S.G. Tzafestas, Ed., New York: Marcel Dekkar, 1986.

[4] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," *IEEE Trans. Auto. Cont.*, vol. AC-32, pp. 115-122, Feb. 1987.

[5] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parameterizations," *Proc. CDC'90*, Honolulu, Dec. 1990, pp. 954-959.

[6] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Sig. Proc.*, vol. 41, pp. 629-637, Feb. 1993.

[7] T. Lin, M. Kawamata, and T. Higuchi, "Minimization of sensitivity of 2-D systems and its relation to 2-D balanced realizations," *Proc. ISCAS'87*, Philadelphia, May 1987, vol. 2, pp. 710-713.

[8] W.S. Lu, E.B. Lee, and Q.T. Zhang, "Model reduction for two-dimensional systems," *Proc. ISCAS'86*, 1986, vol. 1, pp. 79-82.

[9] W.J. Lutz and S.L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circ. Syst.*, vol. 35, pp. 1114-1122, Sept. 1988.

[10] R.H. Middleton and G.C. Goodwin, *Digital Control and Estimation: A Unified Approach*, Englewood Cliffs: Prentice-Hall, 1990.

[11] K. Premaratne, E.I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete-time systems," *IEEE Trans. Circ. Syst.*, vol. CAS-37, pp. 1116-1132, Sept. 1990.

[12] R.P. Roesser, "A discrete state model for linear image processing," *IEEE Trans. Auto. Cont.*, vol. AC-20, pp. 1-10, Feb. 1975.